Trivergence of Probability Distributions, at glance

Juan-Manuel Torres-Moreno[a,b,*]

[a] *Laboratoire Informatique d'Avignon/UAPV BP 91228, 84911 France*
[b] *Ecole Polytechnique de Montréal, Montréal (Québec) Canada*

Abstract

In this paper we introduce the intuitive notion of trivergence of probability distributions (TPD). This notion allow us to calculate the similarity among triplets of objects. For this computation, we can use the well known measures of probability divergences like Kullback-Leibler and Jensen-Shannon. Divergence measures may be used in Information Retrieval tasks as Automatic Text Summarization, Text Classification, among many others.

Keywords: Trivergence of probability distributions, Divergence of probability distributions, Kullback-Leibler Divergence, Jensen-Shannon Divergence

1. Introduction

A statistical distance defines a measure of distance between two objects. This measure of distance may be interpreted as a distance among two probability distributions of two populations. Moreover, a metric is a measure defined on a set $\mathcal{X}$ as a function d such as, $\forall x, y \in \mathcal{X}$, $d : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{R}+$. d respects the following conditions:

i) $d(x, y) \geq 0$
ii) $d(x, y) = 0$ iff $x = y$
iii) $d(x, y) = d(y, x)$
iv) $d(x, z) \leq d(x, y) + d(y, z)$

Several measures of distance are not considered as metrics because they do not fulfill one or more of these conditions. These measures are known as divergences. This is the case of Kullback-Leibler divergence D_{KL}, that in particular, violates the conditions ii) and iii). In other hands, the Jensen-Shannon divergence D_{JS} is a metric. It corresponds to the symmetrical version of the D_{KL} divergence.

[☆] LIA/Université d'Avignon et des Pays de Vaucluse
[*] Corresponding author
Email address: juan-manuel.torres@univ-avignon.fr (Juan-Manuel Torres-Moreno)
URL: lia.univ-avignon.fr/chercheurs/torres (Juan-Manuel Torres-Moreno)

In this paper we introduce the notion of distance among three objects as a trivergence τ of probability distribution. The main idea is based on intuitive properties of divergences.

The rest of the paper is organized as follows: in Section §2 we outline the divergences using probability distributions and smoothing. Section §3 introduces the preliminaries of notion of trivergence. Sections §4 and §5 compute the trivergence as a product of divergences and as a compound divergence function. Finally Section §6 shows the discussion and the conclusions.

2. Preliminaries: divergences of probability distributions with smoothing

In the follows, we recapitulate the divergence functions of probability distributions: the Kullback-Leibler divergence [1] and the Jensen-Shannon symmetrical divergence [2].

2.1. Kullback-Leibler divergence

The divergence of Kullback-Leibler or relative entropy is a distance between two probability distributions p and q is defined by the equation:

$$D_{\mathrm{KL}}(p\|q) = \sum_{w \in p} p_w \log \frac{p_w}{q_w} \tag{1}$$

The logarithm is in base 2, but we adopted the notation convention $\log_2$ as log.

Of course, $q_w = 0$ for a few items w, because not all items of p are in q. In this case, expressions like $p \log \frac{p}{0} \to \infty$ may occur if $q_w = 0$, i.e. when the item $w \notin q$ (see by example the Figure 1). To avoid this situation, in an empirical way, a smoothing process is used for estimating the probability of unseen items. In the literature there are several smoothing techniques, for example Good-Turing, Back-Off, etc. [3, 4]. In this paper, we will use a very elementary smoothing:

$$q_w = \begin{cases} \dfrac{C_w^q}{|q|} & \text{if } w \in q \\ \dfrac{1}{|T|} & \text{elsewhere} \end{cases} \tag{2}$$

where p and q are the probability distributions, $p_w = \frac{C_w^p}{|p|}$, q_w is defined by equation (2), C_w^p is the number of ocurrences of the item $w \in p$, C_w^q is the number of ocurrences of the item $w \in q$, $|p|$ = total number of distinct items $\in p$, $|q|$ = total number of distinct items $\in q$ and $|T| = |p| + |q|$. In other hands, we assume that $|p| > |q|$, then the divergence is calculated from p to q.

The Kullback-Leibler distance is not a metric in a mathematical sense, because despite meeting that $D_{\mathrm{KL}}(p\|q) \geq 0$ with equality if and only if $p = q$, it is not symmetrical and it does not respect the triangle inequality.

2.2. Jensen-Shannon divergence

The Jensen-Shannon divergence[2] or symmetrical distance of Kullback-Leibler between two probability distributions p and q over the same alphabet $\mathcal{X}$ is defined by the equation:

$$D_{\mathrm{JS}}(p\|q) = \frac{1}{2}\left\{\sum_{w\in\mathcal{X}} p_w \log\frac{2p_w}{p_w+q_w} + \sum_{w\in\mathcal{X}} q_w \log\frac{2q_w}{p_w+q_w}\right\} \tag{3}$$

with the same conventions for p, q, $|p|$, $|q|$, $|T|$, p_w, q_w, C_w^p and C_w^q as in equation (1); and the same elementary smoothing (2). The logarithm is also in base 2, but we adopted the same convention for $\log_2$. $\sqrt{D_{\mathrm{JS}}}$ is a metric in a mathematical sense.

3. Trivergence of probability distributions

In order to define the trivergence between three probability distributions we will use divergence measures. Let p, q and r be three probability distributions and $T = \{p \cup q \cup r\}$, with cardinality $|T|$. Figure 1 shows the partitioning of the T set in 7 regions.

We defined two ways to calculate the trivergence τ, as a product of divergences and as a compound divergence function:

i) Product of divergences:

$$\tau^{\pi}(p||q||r) = \begin{cases} D(p||q)\cdot D(q||r)\cdot D(p||r); \\ D(q||p)\cdot D(r||q)\cdot D(r||p) \end{cases} \tag{4}$$

ii) Compound divergence function:

$$\tau^{c}(p||q||r) = \begin{cases} D[\ p||D(q||r)\ ] & ;\ D[\ p||D(r||q)\ ]; \\ D[\ q||D(p||r)\ ] & ;\ D[\ q||D(r||p)\ ]; \\ D[\ r||D(p||q)\ ] & ;\ D[\ r||D(q||p)\ ]; \\ D[\ D(q||r)||p\ ] & ;\ D[\ D(r||q)||p\ ]; \\ D[\ D(p||r)||q\ ] & ;\ D[\ D(r||p)||q\ ]; \\ D[\ D(p||q)||r\ ] & ;\ D[\ D(q||p)||r\ ] \end{cases} \tag{5}$$

In both cases, if we use the following restriction:

$$|p| > |q| > |r|$$

the definition of trivergence is, in particular, sorted by their cardinality. Then, we have for the product:

$$\tau^{\pi}(p||q||r) = D(p||q)\cdot D(q||r)\cdot D(p||r) \tag{6}$$

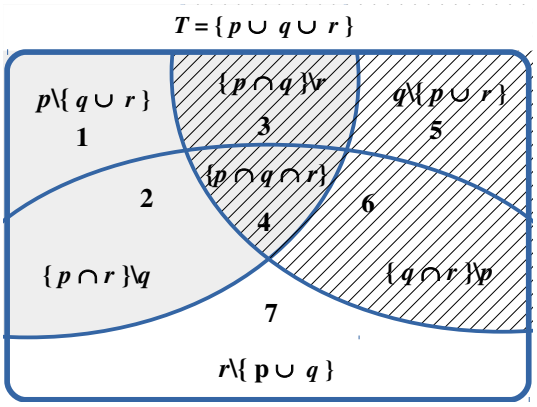

Figure 1: Decomposition of the distributions p, q and r in subsets.

And for the compound function:

$$\tau^c(p||q||r) = D[\ p||D(q||r)\ ] \tag{7}$$

In order to clarify the weight of the smoothing (equation 2) for p_w, q_w and r_w, from Figure 1 we have for each region that:

1. $\{p \backslash \{q \cup r\}\}$: $q_w = r_w = 0$;
2. $\{p \cap r\} \backslash q$: $q_w = 0$;
3. $\{p \cap q\} \backslash r$: $r_w = 0$;
4. $\{p \cap q \cap r\}$: $p_w \neq 0, q_w \neq 0, r_w \neq 0$;
5. $\{q \backslash \{p \cup r\}\}$: $p_w = r_w = 0$;
6. $\{q \cap r\} \backslash p$: $p_w = 0$;
7. $\{r \backslash \{p \cup q\}\}$: $p_w = q_w = 0$.

In this paper, we will use both Kullback-Leibler D_{KL} [1] and Jensen-Shannon D_{JS} [2] divergences in order to calculate the trivergence $\tau^{\pi,c}(p||q||r)$.

4. Distribution using Kullback-Leibler divergence

4.1. τ^π as product of KL divergences

Definition. Let p, q and r be three probability distributions where

$$|p| > |q| > |r|$$

and $T = \{p \cup q \cup r\}$, with cardinality $|T|$. The Kullback-Leibler trivergence between p, q and r, sorted by their cardinality is defined as a product of divergences:

$$\tau^\pi_{\mathrm{KL}}(p||q||r) = D_{\mathrm{KL}}(p||q) \cdot D_{\mathrm{KL}}(q||r) \cdot D_{\mathrm{KL}}(p||r)$$

Calculating simultaneously for p, q and r:

$$D_{\mathrm{KL}}(p||q) = \sum_{x \in p} p_w \log \frac{p_w}{q_w} \tag{8}$$

$$D_{\mathrm{KL}}(q||r) = \sum_{x \in q} q_w \log \frac{q_w}{r_w} \tag{9}$$

$$D_{\mathrm{KL}}(p||r) = \sum_{x \in p} p_w \log \frac{p_w}{r_w} \tag{10}$$

From the equation (8):

$$\sum_{x \in p} p_w \log \frac{p_w}{q_w} \quad = \quad \sum_{x \in p \setminus q} p_w \log \frac{p_w}{q_w} + \sum_{x \in p \cap q} p_w \log \frac{p_w}{q_w} \tag{11}$$

and using the smoothing from the equation (2):

$$\sum_{x \in p} p_w \log \frac{p_w}{q_w} = \begin{cases} \displaystyle\sum_{x \in p \setminus q} \frac{C_w^p}{|p|} \log \frac{|T| C_w^p}{|p|} & \text{smooth } q_w = \dfrac{1}{|T|} \\ \displaystyle\sum_{x \in p \cap q} \frac{C_w^p}{|p|} \log \frac{|q|}{|p|} \frac{C_w^p}{C_w^q} & \text{without smooth} \end{cases}$$

From the equation (9):

$$\sum_{x \in q} q_w \log \frac{q_w}{r_w} \quad = \quad \sum_{x \in q \setminus r} q_w \log \frac{q_w}{r_w} + \sum_{x \in q \cap r} q_w \log \frac{q_w}{r_w} \tag{12}$$

and using the smoothing from the equation (2):

$$\sum_{x \in q} q_w \log \frac{q_w}{r_w} = \begin{cases} \displaystyle\sum_{x \in q \setminus r} \frac{C_w^q}{|q|} \log \frac{|T| C_w^q}{|q|} & \text{smooth } r_w = \dfrac{1}{|T|} \\ \displaystyle\sum_{x \in \{q \cap r\}} \frac{C_w^q}{|q|} \log \frac{|r|}{|q|} \frac{C_w^q}{C_w^r} & \text{without smooth} \end{cases}$$

From the equation (10):

$$\sum_{x \in p} p_w \log \frac{p_w}{r_w} \quad = \quad \sum_{x \in p \setminus r} p_w \log \frac{p_w}{r_w} + \sum_{x \in p \cap r} p_w \log \frac{p_w}{r_w} \tag{13}$$

and using the smoothing from the equation (2):

$$\sum_{x \in p} p_w \log \frac{p_w}{r_w} = \begin{cases} \displaystyle\sum_{x \in p \setminus r} \frac{C_w^p}{|p|} \log \frac{|T| C_w^p}{|p|} & \text{smooth } r_w = \dfrac{1}{|T|} \\ \displaystyle\sum_{x \in \{p \cap r\}} \frac{C_w^p}{|p|} \log \frac{|r|}{|p|} \frac{C_w^p}{C_w^r} & \text{without smooth} \end{cases}$$

therefore:

$$D_{\mathrm{KL}}(p||q) = \sum_{x \in p \setminus q} \frac{C_w^p}{|p|} \log \frac{|T| C_w^p}{|p|} + \sum_{x \in \{p \cap q\}} \frac{C_w^p}{|p|} \log \frac{|q|}{|p|} \frac{C_w^p}{C_w^q} \tag{14}$$

$$D_{\mathrm{KL}}(q||r) = \sum_{x \in q \setminus r} \frac{C_w^q}{|q|} \log \frac{|T| C_w^q}{|q|} + \sum_{x \in \{q \cap r\}} \frac{C_w^q}{|q|} \log \frac{|r|}{|q|} \frac{C_w^q}{C_w^r} \tag{15}$$

$$D_{\mathrm{KL}}(p||r) = \sum_{x \in p \setminus r} \frac{C_w^p}{|p|} \log \frac{|T| C_w^p}{|p|} + \sum_{x \in \{p \cap r\}} \frac{C_w^p}{|p|} \log \frac{|r|}{|p|} \frac{C_w^p}{C_w^r} \tag{16}$$

4.2. τ^π *as compound divergence function*

Definition Let p, q and r be three probability distributions where

$$|p| > |q| > |r|$$

and $T = \{p \cup q \cup r\}$, with cardinality $|T|$. The Kullback-Leibler trivergence between p, q and r, sorted by their cardinality is defined as a compound divergence function:

$$\tau_{\mathrm{KL}}^c(p||q||r) = D_{\mathrm{KL}} \left[\; p \; || \; \frac{D_{\mathrm{KL}}(q||r)}{|q|} \; \right]$$

We computed $\frac{D_{\mathrm{KL}}(q||r)}{|q|}$ in order to consider this fraction suchs as a probability.

Firstly, we calculate:

$$D_{\mathrm{KL}}(q||r) \quad = \quad \sum_{w \in q} q_w \log \frac{q_w}{r_w}$$

however $\sum_{w \in q} q_w \log \frac{q_w}{r_w}$ is defined by equation (15), therefore using a smoothing in the case of unseen events:

$$\tau_{\mathrm{KL}}^c(p||q||r) = \begin{cases} \displaystyle\sum_{x \in p \cap q} p_x \log \frac{|q| p_x}{D_{\mathrm{KL}}(q||r)} & \\ \displaystyle\sum_{x \in p \setminus q} p_x \log |T| p_x & \text{if } D_{\mathrm{KL}}(q||r) = 0; \end{cases} \tag{17}$$

5. Distribution using Jensen-Shannon divergence

5.1. τ^π *as product of JS divergences*

Definition. Let p, q and r be three probability distributions where

$$|p| > |q| > |r|$$

and $T = \{p \cup q \cup r\}$, with cardinality $|T|$. The Jensen-Shannon trivergence between p, q and r, sorted by their cardinality is defined as a product of divergences:

$$\tau_{\mathrm{JS}}^\pi(p||q||r) = D_{\mathrm{JS}}(p||q) \cdot D_{\mathrm{JS}}(q||r) \cdot D_{\mathrm{JS}}(p||r)$$

We defined:

$$P_w^{pq} = p_w \log \frac{2p_w}{p_w + q_w}; Q_w^{pq} = q_w \log \frac{2q_w}{p_w + q_w}$$

$$Q_w^{qr} = q_w \log \frac{2q_w}{q_w + r_w}; R_w^{qr} = r_w \log \frac{2r_w}{q_w + r_w}$$

$$R_w^{pr} = r_w \log \frac{2r_w}{r_w + p_w}; P_w^{pr} = p_w \log \frac{2p_w}{r_w + p_w}$$

Calculating simultaneously for p, q and r:

$$D_{\mathrm{JS}}(p||q) = \frac{1}{2} \sum_{w \in \{p \cup q\}} \{P_w^{pq} + Q_w^{pq}\} \tag{18}$$

$$D_{\mathrm{JS}}(q||r) = \frac{1}{2} \sum_{w \in \{q \cup r\}} \{Q_w^{qr} + R_w^{qr}\} \tag{19}$$

$$D_{\mathrm{JS}}(p||r) = \frac{1}{2} \sum_{w \in \{p \cup r\}} \{P_w^{pr} + R_w^{pr}\} \tag{20}$$

For $2D_{\mathrm{JS}}(p||q)$ we have:

$$\begin{aligned} \sum_{w \in p \cup q} \{P_w^{pq} + Q_w^{pq}\} &= \sum_{w \in p \setminus q} P_w^{pq} + Q_w^{pq} + \sum_{w \in p \cap q} P_w^{pq} + Q_w^{pq} \\ &+ \sum_{w \in q \setminus p} P_w^{pq} + Q_w^{pq} \end{aligned}$$

and using the smoothing for p_w and q_w from the equation (2):

$$\sum_{w \in p \cup q} P_w^{pq} + Q_w^{pq} = \begin{cases} \sum_{w \in p \setminus q} \frac{C_w^p}{|p|} \log \frac{2|T|C_w^p}{|T|C_w^p + |p|} + \frac{1}{T} \log \frac{2|p|}{|T|C_w^p + |p|};\ q_w = \frac{1}{|T|} \\ \sum_{w \in p \cap r} \frac{C_w^p}{|p|} \log \frac{2|q|C_w^p}{|q|C_w^p + |p|C_w^q} + \frac{C_w^q}{|q|} \log \frac{2|p|C_w^q}{|q|C_w^p + |p|C_w^q} \\ \sum_{w \in q \setminus p} \frac{1}{T} \log \frac{2|q|}{|T|C_w^q + |q|} + \frac{C_w^q}{|q|} \log \frac{2|T|C_w^q}{|T|C_w^q + |q|};\ p_w = \frac{1}{|T|} \end{cases} \tag{21}$$

For $2D_{\mathrm{JS}}(q||r)$ we have:

$$\begin{aligned} \sum_{w \in q \cup r} \{Q_w^{qr} + R_w^{qr}\} &= \sum_{w \in q \setminus r} Q_w^{qr} + R_w^{qr} + \sum_{w \in q \cap r} Q_w^{qr} + R_w^{qr} \\ &+ \sum_{w \in r \setminus q} Q_w^{qr} + R_w^{qr} \end{aligned}$$

and using the smoothing for q_w and r_w from the equation (2):

$$\sum_{w\in q\cup r} Q_w^{qr} + R_w^{qr} = \begin{cases} \sum_{w\in q\setminus r} \frac{C_w^q}{|q|}\log\frac{2|T|C_w^q}{|T|C_w^q+|q|} + \frac{1}{T}\log\frac{2|q|}{|T|C_w^q+|q|};\ r_w=\frac{1}{|T|} \\ \sum_{w\in q\cap r} \frac{C_w^q}{|q|}\log\frac{2|r|C_w^q}{|r|C_w^q+|q|C_w^r} + \frac{C_w^r}{|r|}\log\frac{2|q|C_w^r}{|r|C_w^q+|q|C_w^r} \\ \sum_{w\in r\setminus q} \frac{1}{T}\log\frac{2|r|}{|T|C_w^r+|r|} + \frac{C_w^r}{|r|}\log\frac{2|T|C_w^r}{|T|C_w^r+|r|};\ q_w=\frac{1}{|T|} \end{cases} \tag{22}$$

Finally, for $2D_{\mathrm{JS}}(p||r)$ we have:

$$\begin{aligned} \sum_{w\in p\cup r}\{P_w^{pr}+R_w^{pr}\} &= \sum_{w\in p\setminus r} P_w^{pr}+R_w^{pr} + \sum_{w\in p\cap r} P_w^{pr}+R_w^{pr} \\ &+ \sum_{w\in r\setminus p} P_w^{pr}+R_w^{pr} \end{aligned}$$

Using the smoothing for p_w and r_w from the equation (2):

$$\sum_{w\in p\cup r} P_w^{pr} + R_w^{pr} = \begin{cases} \sum_{w\in p\setminus r} \frac{C_w^p}{|p|}\log\frac{2|T|C_w^p}{|T|C_w^p+|p|} + \frac{1}{T}\log\frac{2|p|}{|T|C_w^p+|p|};\ r_w=\frac{1}{|T|} \\ \sum_{w\in p\cap r} \frac{C_w^p}{|p|}\log\frac{2|r|C_w^p}{|r|C_w^p+|p|C_w^r} + \frac{C_w^r}{|q|}\log\frac{2|p|C_w^q}{|r|C_w^p+|p|C_w^r} \\ \sum_{w\in r\setminus p} \frac{1}{T}\log\frac{2|r|}{|T|C_w^r+|r|} + \frac{C_w^r}{|r|}\log\frac{2|T|C_w^r}{|T|C_w^r+|r|};\ p_w=\frac{1}{|T|} \end{cases} \tag{23}$$

5.2. τ^c as compound divergence function

Definition. Let p, q and r be three probability distributions where

$$|p| > |q| > |r|$$

$T = \{p \cup q \cup r\}$, with cardinality $|T|$ and $QR = \{q \cup r\}$, with cardinality $|QR|$. The Jensen-Shannon trivergence sorted by their cardinality, between p, q and r is defined as a compound divergence function:

$$\tau_{\mathrm{JS}}^c(p||q||r) = D_{\mathrm{JS}}\left[\ p\ ||\ \frac{D_{\mathrm{JS}}(q||r)}{|QR|}\ \right]$$

We computed $\frac{D_{\mathrm{JS}}(q||r)}{|q|+|r|}$ in order to consider this fraction suchs as a probability.

First, we calculate:

$$D_{\mathrm{JS}}(q||r) = \frac{1}{2}\left\{\sum_{w\in q\cup r} q_w\log\frac{2q_w}{q_w+r_w} + \sum_{w\in q\cup r} r_w\log\frac{2r_w}{q_w+r_w}\right\}$$

neverthless $D_{\mathrm{JS}}(q||r)$ is defined by equation (22), therefore using a smoothing in the case of unseen events:

$$p_x = \frac{D_{\mathrm{JS}}(q||r)}{|q| + |r|} = \frac{1}{|T|}$$

$$\tau^{c}_{\mathrm{JS}}(p||q||r) = \frac{1}{2} \times$$

$$\begin{cases} \displaystyle\sum_{x \in p \cap \{q \cup r\}} p_x \log \frac{2|QR|p_x}{|QR|p_x + D_{\mathrm{JS}}(q||r)} + \frac{D_{\mathrm{JS}}(q||r)}{|QR|} \log \frac{2D_{\mathrm{JS}}(q||r)}{|QR|p_x + D_{\mathrm{JS}}(q||r)} \\ \displaystyle\sum_{x \in p \setminus \{q \cup r\}} p_x \log \frac{2|T|p_x}{|T|p_x + 1} + \frac{1}{|T|} \log \frac{2}{|T|p_x + 1} & ; \text{if } D_{\mathrm{JS}}(q||r) = 0 \\ \displaystyle\sum_{x \in \{q \cup r\} \setminus p} \frac{1}{|T|} \log \frac{2|QR|}{|QR| + |T|D_{\mathrm{JS}}(q||r)} + \frac{D_{\mathrm{JS}}(q||r)}{|QR|} \log \frac{2|T|D_{\mathrm{JS}}(q||r)}{|QR| + |T|D_{\mathrm{JS}}(q||r)} \\ & ; \text{if } p_x = 0 \end{cases}$$

6. Conclusions

The main contribution of this paper is the formalisation of the definition of smoothed Trivergence of Probability Distributions (TPD). The trivergence of three objects represented as probability distributions, was calculated using elementary functions of divergence (KL and JS). We have proposed two ways to compute the smoothed TPD. The first one uses a product of divergences and the second one uses a compound divergence function. Divergences measures hase been used in Automatic Text Summarization [5, 6, 7] tasks among many others.

References

References

[1] T. M. Cover, J. A. Thomas, Elements of information theory, Wiley, New York, 1991.

[2] D. M. Endres, J. E. Schindelin, A New Metric for Probability Distributions, IEEE Trans. Inform. Theory 49 (7) (2003) 1858–1860.

[3] S. F. Chen, An empirical study of smoothing techniques for language modeling, Tech. rep. (1998).

[4] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, 1999.

[5] A. Louis, A. Nenkova, Automatically Assessing Machine Summary Content Without a Gold Standard, Computational Linguistics 39 (2) (2013) 267–300.

[6] J.-M. Torres-Moreno, H. Saggion, I. da Cunha, E. SanJuan, P. Velazquez-Morales, Summary Evaluation With and Without References, Polibits: Research Journal on Computer Science and Computer Engineering with Applications 42 (2010) 13–19.

[7] H. Saggion, J.-M. Torres-Moreno, I. da Cunha, E. SanJuan, P. Velazquez Morales, Multilingual Summarization Evaluation Without Human Models, in: 23rd COLING International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 1059–1067.
URL aclweb.org/anthology/C/C10/C10-2122.pdf